Australian Bureau of Statistics

6102.0 - Labour Statistics: Concepts, Sources and Methods, 2001

ARCHIVED ISSUE Released at 11:30 AM (CANBERRA TIME) 18/05/2001 Ceased

Contents >> Methods >> Chapter 16. Overview of Survey Methods

INTRODUCTION

16.1 The methods part of this publication describes the major ABS statistical surveys in the field of labour statistics and their compilation methods. Detailed information on scope, coverage, sample design, collection processes, estimation techniques and statistical output is presented for each of the surveys.

16.2 This chapter provides an overview of key aspects of survey design. It defines and explains key concepts and terms that relate to survey design. It commences with a discussion of sample surveys and censuses, and collection methodologies used in ABS surveys. The rest of the chapter is organised into the following topics: sample design and sampling techniques; estimation; time series estimates; reliability of estimates; output; and data comparability over time.

16.3 The subsequent chapters are organised into two sections. ABS household surveys are presented in chapters 17 to 22, and ABS business surveys in chapters 23 to 31. Each section begins with a chapter outlining aspects of survey methodology which are common to the type of survey being discussed (i.e. household or business survey). A separate chapter is then devoted to each major labour-related ABS survey.

SAMPLE SURVEYS VERSUS CENSUSES

16.4 The ABS uses both sample surveys and censuses to collect information from a population about characteristics of interest. In the field of labour statistics, the ABS uses sample surveys of both households and businesses, as well as censuses (such as the Industrial Disputes collection).

16.5 Censuses involve the collection of information from all units in the target population, while sample surveys involve the collection of information from only a part (sample) of the target population.

16.6 Sample surveys have both advantages and disadvantages when compared with censuses. Some advantages are reduced costs (as less time is needed to collect, process and produce data), possible reductions in non-sampling error (this concept is discussed in further detail later in this chapter), improved timeliness, and the potential to gather more detailed information from each respondent.

16.7 A disadvantage of sample surveys is that estimates are subject to sampling error, which occurs because data were obtained from only a sample rather than the entire population (this concept is discussed in further detail later in this chapter). Also, as a result of obtaining only a small number of observations in particular geographical areas and sub-populations, detailed cross-tabulations may be subject to high levels of error, and hence of limited use.

16.8 Censuses are generally used when accurate information is sought for many sub-groups of the population. Collecting this type of information from a sample survey would require a very large sample.

COLLECTION METHODOLOGY

16.9 A number of methods are used by the ABS for collecting data. Those most commonly used in labour-related surveys can be categorised into three basic groups:

- interview;
- self-enumeration; and
- documentary sources.

INTERVIEW

16.10 The interview method of data collection involves an interviewer contacting data providers, asking the questions, and recording the responses. Interviews can be either personal or involving Any Responsible Adult (ARA), and can be conducted either face to face or over the telephone. Interviews are most commonly used in household surveys.

Personal interview

16.11 Personal interviewing involves each provider being questioned about his or her own details.

Any Responsible Adult interview

16.12 The Any Responsible Adult (ARA), or proxy, method of interviewing is used in a number of ABS household surveys as an alternative to personal interviewing. This involves obtaining information about all the persons in a selected household who are in scope of the survey, from the first responsible adult with whom the interviewer makes contact (rather than speaking to each individual personally). The method is only used for collecting information on topics where other members of the household are likely to be able to answer the question. If the ARA is unable to supply all of the details for another individual in the household, a personal interview is conducted.

Face to face interview

16.13 Face to face interviews involve a trained interviewer visiting the provider to conduct the survey. Advantages of this method of data collection are higher response rates and improved data quality. Interviewers are able to help respondents understand the questions and provide correct answers, thereby allowing for the collection of more complex data. The improved quality of the data means that less data editing and correction is required at a later stage.

16.14 However, face to face interviews are expensive. There are costs involved in time and travel to reach the respondents, and in the recruitment, training, and management of an interviewer work force. Other disadvantages are that data can possibly be subject to bias caused by the interviewer's appearance and attitude, and that respondents may not feel free to disclose sensitive or private information to an interviewer.

Telephone interview

16.15 In telephone interviews the providers are asked the survey questions over the telephone. This reduces the costs compared to face to face interviews as fewer interviewers are needed and there are no travel costs involved. Telephone interviews can also produce more timely results. Call-backs for 'not-answering' and follow-ups for additional information are relatively quick and inexpensive.

16.16 As with other methods of data collection, there are some drawbacks associated with this approach. There are limits on the number and complexity of questions that can be asked and, because of the ease with which the respondent can terminate the interview, non-response and partial non-response can be higher than with face to face interviews.

16.17 Telephone interviewing is used in both ABS household and business surveys, sometimes in conjunction with face to face interviews. For example, in the Labour Force Survey the first interview is conducted face to face. The remaining interviews are conducted by telephone if the provider agrees.

Computer-assisted interview

16.18 When performing a computer-assisted interview, the interviewer enters the data into a computer as they are provided. This allows some edit checks to be carried out at the time of the interview, improving data quality. Overall timeliness of data processing is also improved.

SELF-ENUMERATION

16.19 Self-enumeration surveys are those in which it is left to the providers to complete the survey questionnaires. Three of the most common self-enumeration methods are: dropoff/mail-back, dropoff/pickup, and mail-out/mail-back. These are discussed below. Self-enumeration surveys are most commonly used in business surveys, but can be used in household surveys.

Dropoff/Mail-back and Dropoff/Pickup

16.20 Dropoff/mail-back and dropoff/pickup methods are used in a number of ABS household surveys. These two closely related methods of self-enumeration both provide higher response rates and data quality than postal surveys. In both cases the questionnaire is delivered to respondents by an interviewer who explains the aims of the survey and how to fill out the questionnaire. The questionnaire is left with the respondent to be completed, and either mailed back or picked up at a later date. An example of a dropoff/pickup survey is the Census of Population and Housing.

Mail-out/mail-back

16.21 Mail-out/mail-back surveys are used most commonly in ABS business surveys. This approach involves mailing questionnaires to respondents with a return-paid envelope so that the respondent can mail back the completed questionnaire. It allows wide geographic areas to be covered at a relatively low cost (compared to personal interviews), and allows access to 'difficult-to-contact' respondents (e.g. where a post office box is the only address provided). It also allows respondents to complete questionnaires in their own time. Another advantage of mail-out/mail-back surveys is that respondents may feel more comfortable providing data directly to the ABS without divulging confidential information to an interviewer. There are some disadvantages with this approach: response rates can be low; there can be delays between the time the questionnaire is sent out and returned; there are limits on the length and complexity of the questionnaire; and it is necessary to have a complete list of addresses for all units in the sample.

DOCUMENTARY SOURCES (ADMINISTRATIVE DATA)

16.22 This method involves the use of existing data, such as administrative data, to obtain information about the survey population as a whole and about individual units. The approach is used in some ABS business surveys to collect information about individual units in the survey population. Payroll records from some government departments, for example, are used in business surveys that collect information on earnings and employment.

16.23 An advantage of using administrative data is that it can save both time and money by removing the need for the ABS to collect the information directly from respondents. Disadvantages of using administrative data are that: often the data quality is below ABS quality standards, requiring substantial manipulation and checking before the data can be used (adding to the expense); the underlying concepts relate to administrative procedures rather than statistical constructs; and sometimes not all the data required for statistical purposes have been collected, or they have not been collected in a manner suitable for the purposes of the ABS.

SAMPLE DESIGN AND SAMPLING TECHNIQUES

16.24 All the ABS labour-related sample surveys referred to in this publication (household and business) use probability sampling techniques, drawing their samples from a population frame. This section briefly defines and explains key concepts and terms related to survey design. Subsequent chapters provide more detail on aspects of survey design that are particular to household surveys (Chapter 17) and business surveys (Chapter 23).

POPULATION

16.25 A survey is concerned with two types of population: the target population, and the survey population. The **target population** is the group of units about which information is sought and is also known as the **scope** of the survey. It is the population at which the survey is aimed. The scope should state clearly the units from which data are required, and the extent and time covered e.g. households (units) in Australia (extent) in August 2000 (time).

16.26 However, the target population is a theoretical population. In practice, there are usually a number of units in the target population which cannot be surveyed. These include units which are difficult to contact and units which are missing from the frame (see 16.28). The **survey population** is that part of the population that is able to be surveyed; it is also called the **coverage population**.

STATISTICAL UNITS

16.27 Statistical units are used in the design, collection, analysis and dissemination of statistical data. There are several types of units, including: sampling units (the units selected in the sample survey), collection units (the units from which data are collected), reporting units (the units about which data are collected), and analysis units (the units used for analysis of the data). The units used in a survey may change at various stages in the survey cycle. For example, the Labour Force Survey uses a sample of dwellings (sampling unit) from which information is collected from any responsible adult (collection unit) about each person in the household in scope of the survey (reporting units). The results of the survey may then be analysed for families (analysis unit).

FRAME

16.28 The frame comprises a list of statistical units (e.g. persons, households or businesses) in the population, together with auxiliary information about each unit. It serves as a basis for selecting the sample. Two types of frames are used in ABS labour-related surveys: list based frames and area based frames.

List based frames

16.29 List based frames comprise a list of all sampling units in the survey population. List based frames are commonly used in surveys of businesses. ABS business surveys currently draw their list frames from the ABS Business Register. The ABS Business Register is discussed further in Chapter 23.

Area based frames

16.30 Area based frames comprise a list of non-overlapping geographic areas. These areas may be defined by geographical features such as rivers and streets. They are usually used in household surveys. Once an area is selected, a list is made of the dwellings in the area, and a sample of dwellings selected from the list. An area based frame obviates the need to maintain a complete listing of all dwellings in Australia, leading to cost savings. Examples of geographic areas that may be used to create area frames include: local government areas; census collection districts; and postcodes.

Auxiliary variables

16.31 Auxiliary variables are characteristics of each unit for which information is known on the frame prior to the survey. Auxiliary variables can be used in the sample design to better target the population of interest, if the information on the frame is of sufficiently high quality. They can also be used in the estimation process in conjunction with the survey data.

Frame problems

16.32 For most sampling methodologies, it is desirable to have a complete list from which to select a sample. However, in practice it can be difficult to compile such a complete list and therefore **frame bias** may be introduced. Frame bias occurs when an inappropriate frame is used or there are problems with the composition of the frame, with the result that the frame is not representative of the target population. Frames become inaccurate for many reasons. One of the most common problems is that populations change continuously, causing frames to become out of date. Frames may also be inaccurate if they are compiled from inaccurate sources. The following are some of the problems that can occur in the composition of frames.

16.33 **Undercoverage** occurs when some units in the target population that should appear on the frame do not. These units may have different characteristics from those units which appear on the frame, and therefore results from the survey will not be representative of the target population.

16.34 **Out of scope units** are units that appear on the frame but are not elements of the target population. Selection of a number of out of scope units in the sample reduces the effective sample size, and increases standard errors. Furthermore, out of scope units appearing on the frame may be incorrectly accounted for in the estimation process which may lead to bias in survey estimates.

16.35 **Duplicates** are units that appear more than once on the frame. The occurrence of duplicates means that the probability of selection of the units on the frame is no longer known. In particular, the duplicate units will have more than the correct chance of selection, introducing bias towards the characteristics of these units. Duplicates also increase standard errors.

16.36 **Deaths** are units that no longer exist in the population but are still on the frame. Deaths have the same impact on survey results as out of scope units.

16.37 The **quality of auxiliary variables** can affect the survey estimates of the variables of interest through both the survey design and the estimation process.

16.38 The ABS attempts to minimise frame problems. The ABS uses standardised sample and frame maintenance procedures across collections. Some of the approaches taken are to adjust estimates using new business provisions (explained further in Chapter 23), and to standardise across surveys the systems for handling estimation, imputation and outliers (explained in Estimation and Weighting).

PROBABILITY SAMPLES

16.39 Probability samples are samples drawn from populations, such that every unit in the population has a known, or calculable, non-zero probability of selection which can be obtained prior to selection. In order to calculate the probability of selection, a population frame must be available. The sample is then drawn from this frame. Alternatives to probability samples are samples formed without a frame, such as phone-in polls.

16.40 Probability sampling is the preferred ABS method of conducting major surveys, especially when a population frame is available. Probability samples allow estimates of the accuracy of the survey estimates to be calculated. They are also used in ABS surveys as a means of avoiding bias in survey results. Bias is avoided when either the probability of selection is equal for all units in the population or, where this is not the case, the effect of non-equal probabilities is allowed for in estimation.

STRATIFIED SAMPLING

16.41 Stratified sampling is a technique which uses auxiliary information available for every unit on the frame to increase the efficiency of a sample design. Stratified sampling involves the division or stratification of the population frame into homogeneous (similar) groups called strata, which can be treated as totally separate populations. A sample is then selected independently from each of these groups, and can therefore be selected in different ways for different strata, e.g. some strata may be sampled using 'simple random sampling' while others may be 'completely enumerated'. These terms are explained below. Stratification variables may be geographical (e.g. State, capital city/balance of State) or non-geographical (e.g. number of employees, industry, turnover).

16.42 All surveys conducted by the ABS use stratification. Household surveys use mainly geographic strata. Business surveys typically use strata which are related to the economic activity undertaken by the business, for example industry and size of the business (the latter based on employment size).

Completely enumerated strata

16.43 Completely enumerated (CE) strata are strata in which information is obtained from all units. Strata that are completely enumerated tend to be those where: each population unit within the stratum is likely to contribute significantly to the estimate being produced (such as strata containing large employers where the estimate being produced is employment); or there is significant variability across the population units within the stratum.

SIMPLE RANDOM SAMPLING

16.44 Simple random sampling is a probability sampling scheme in which each possible sample of the required size has the same chance of selection. It follows that each unit of the population has an equal chance of selection.

16.45 Simple random sampling can involve units being selected either with or without replacement. Replacement sampling allows the units to be selected multiple times, whereas without replacement sampling allows a unit to be selected only once. In general, simple random sampling without replacement produces more accurate results as it does not allow sample to be 'wasted' on duplicate selections. All ABS surveys that use simple random sampling use the 'without replacement' variant. Simple random sampling without replacement is used in most ABS business surveys.

SYSTEMATIC SAMPLING

16.46 Systematic sampling is used in most ABS household surveys, and provides a simple method of selecting the sample. It involves choosing a random starting point within the frame and then applying a fixed interval (referred to as the 'skip') to select members from a frame.

16.47 Information on auxiliary variables can be used in systematic sampling to improve the efficiency of the sample. The units in the frame can be ordered with respect to auxiliary variables prior to calculating the skip interval and starting point. This approach ensures that the sample is spread throughout the range of units on the frame, ensuring a more representative sample.

16.48 Systematic sampling with ordering by auxiliary variables is only useful if the frame contains auxiliary variables about each of the units in the population, and if these variables are related to the variables of interest. The relationship between the variables of interest and the auxiliary variables is often not uniform across strata. Consequently it is possible to design a sample survey with only some of the strata making use of auxiliary variables.

PROBABILITY PROPORTIONAL TO SIZE SAMPLING

16.49 Probability proportional to size sampling is a selection scheme in which units in the population do not all have the same chance of selection. With this method, the larger the unit with respect to some measure of size, the greater the probability that unit will be selected in the sample. Probability proportional to size sampling will lead to unbiased estimates, provided the different probabilities of selection are accounted for in estimation.

CLUSTER SAMPLING

16.50 Cluster sampling involves the units in the population being grouped into convenient clusters, usually occurring naturally. These clusters are non-overlapping, well-defined groups which usually represent geographical areas. The sample is selected by selecting a number of clusters, rather than directly selecting units. All units in a selected cluster are included in the sample.

MULTI-STAGE SAMPLING

16.51 Multi-stage sampling is an extension of cluster sampling. It involves selecting a sample of clusters (first-stage sample) and then selecting a sample of population units within each selected cluster (second-stage sample). The sampling unit changes at each stage of selection. Any number of stages can be employed. The sampling units for any given stage of selection each form clusters of the next-stage sampling units. Units selected in the final stage of sampling are called final-stage units (or ultimate sampling units). The Survey of Employee Earnings and Hours uses multi-stage sampling - businesses (the first-stage units) selected in the survey are asked to select a sample of 'employees' (the final-stage units) using employee payrolls. Household surveys also use multi-stage sampling.

MULTI-PHASE SAMPLING

16.52 Multi-phase sampling involves collecting basic information from a sample of population units, then taking a sub-sample of these units (the second-phase sample) to collect more detailed information. The second-phase sample is selected using the basic information supplied, and allows the second-phase sample to be targeted to the specific population of interest. Population totals for auxiliary variables, and values from the first-phase sample are used to weight the second-phase sample for the estimation of population totals.

16.53 Multi-phase sampling aims to reduce sample size (and hence respondent burden and collection costs) while ensuring that a representative sample is still selected from the population of interest. It is often used when the population of interest is small and difficult to isolate in advance, or when detailed information is required. Multi-phase sampling is also useful when auxiliary information is not known for all of the frame units, as it enables the collection of data for auxiliary variables in the first-phase sample.

16.54 The first-phase sample is designed to be large to ensure sufficient coverage of the population of interest, but only basic information is collected. The basic information is then used to identify those first-phase sample units which are part of the population of interest. A sample of these units is then selected for the second-phase sample. Therefore the sampling unit remains the same for each phase of selection. If multi-phase sampling was not used, detailed information would need to be collected from all of the first-phase sample units to ensure reasonable survey estimates. Therefore, multi-phase sampling reduces the overall respondent burden.

ESTIMATION

16.55 Sample survey data only relate to the units in the sample. Therefore the sample estimates need to be inflated to represent the whole population of interest. Estimation is the means by which this inflation occurs.

16.56 The following section outlines various methods of calculating the population estimates from the sample survey data. It then describes various editing procedures used in labour-related statistics to improve the population estimates.

WEIGHTING

16.57 Estimation is essentially the application of weights to the individual survey records. The value of these weights is determined with respect to one or more of the following three factors:

- the probability of selection for each survey unit (probability weighting);

- adjustment for non-response to correct for imbalances in the characteristics of responding sample units (post-stratification); and
- adjustments to agree with known population totals for auxiliary variables - to correct for further imbalances in the characteristics of the selected sampled units (post-stratification, ratio estimation, calibration).

16.58 Weights are determined using formulae (estimators) of varying complexity.

NUMBER-RAISED ESTIMATION

16.59 Number-raised weights are given by N/n (where N is the total number of units in the population for the stratum, and n is the number of responding units in the sample for that stratum). The weight assigned to each survey unit indicates the number of units in the target population that the survey unit is meant to represent. For example, a survey unit with a weight of 100 represents 100 units in the population. Using number-raised weights, each survey unit in a stratum is given the same weight. Number-raised weights can only be used to weight simple random samples.

16.60 Advantages of number-raised estimation are: it does not require auxiliary data; it is unbiased; and the accuracy of the estimates can be calculated relatively simply. However, number-raised estimation is not as accurate as some other methods.

RATIO ESTIMATION

16.61 Ratio estimation involves the use of known population totals for auxiliary variables to improve the weighting from sample values to population estimates. It operates by comparing the survey sample estimate for an auxiliary variable with the known population total for the same variable on the frame. The ratio of the sample estimate of the auxiliary variable to its population total on the frame is used to adjust the sample estimate for the variable of interest.

16.62 The ratio weights are given by X/x (where X is the known population total for the auxiliary variable, and x is the corresponding estimate of the total based on all responding units in the sample). These weights assume that the population total for the variable of interest will be estimated by the sample equally as well (or poorly) as the population total for the auxiliary variable is estimated by the sample.

16.63 Ratio estimation can be more accurate than number-raised estimation if the auxiliary variable is highly correlated with the variable of interest. However it is slightly biased, with the bias increasing for smaller sample sizes and where there is lower correlation between the auxiliary variable and the variable of interest.

POST-STRATIFICATION

16.64 Post-stratification estimation also involves the use of auxiliary information to improve the weighting from sample values to population estimates. Subgroups of the survey sample units are formed, based on auxiliary variables, after the survey data have been collected. Estimates of subgroup population sizes (based on probability weighting) are compared with known subgroup population sizes from independent sources. The ratio of the two population sizes for each subgroup is used to adjust the original estimate for the variable of interest (based on probability sampling).

16.65 Post-stratification is used to refine the estimation weighting process by correcting for sample imbalance and, assuming that the survey respondents are representative of missing units, correcting for non-response. For example, in the Labour Force Survey, the sample is post-stratified by age, sex, capital city/rest of State, and State/Territory of usual residence. Estimates of the number of people in these subgroups based on Census data are then compared to the estimates based on the survey sample to give the post-stratification weights.

CALIBRATION

16.66 Calibration essentially uses all available auxiliary information to iteratively modify the original weights (based on number-raised weights). The new weights ensure that the sample estimates are consistent with the various auxiliary information. Both post-stratification and ratio estimation can be used as part of the calibration weighting process. Calibration is useful if the survey sample estimates need to match the unit totals for a number of different subgroups or for more that one auxiliary variable. It is mostly used in Special Social Surveys. For example, the Survey of Employment and Unemployment Patterns was weighted so that the survey estimates aligned with both population estimates based on Census data and estimates of the number of people 'employed', 'unemployed' and 'not in the labour force' from the Labour Force Survey.

EDITING

16.67 Editing is the process of correcting data suspected of being wrong, in order to allow the production of reliable statistics. Editing occurs prior to weighting and should aim:

- to ensure that outputs from the collection are mutually consistent: for example, two different methods of deriving the same value should give the same answer;
- to correct for any missing data;
- to detect major errors, which could have a significant effect on the outputs; and
- to find any unusual output values and their causes.

16.68 The purpose of editing is to correct non-sampling errors such as those introduced by misunderstanding of questions or instructions, interviewer bias, miscoding, non-availability of data, incorrect transcription, non-response, and non-contact. Non-response occurs when all (total non-response) or part (partial non-response) of a questionnaire is not completed by the respondent. Non-response is a serious problem and can cause bias in the sample based estimates.

16.69 Editing is also used to identify outliers. The statistical term 'outlier' has several definitions depending on the context in which it is used. Here it is used loosely to describe extreme values that are verified as being correct, but are very different from the values reported by similar units, and are expected to occur only very rarely in the population as a whole. In practice, an outlier is usually considered to be a unit that has a large effect on survey estimates of level, on estimates of movement, or on the sampling variance. This may occur because the unit is not similar to other units in the stratum - for example, if its true employment is much greater than the frame employment. It may also occur when an extreme value is recorded for some variable from an otherwise ordinary sampling unit. The presence of outliers in the sample, particularly in strata with small sampling fractions[1], may result in grossly inadequate estimates, unless they are treated in a special way.

[1] The sampling fraction for a stratum is defined as n/N (where n is the number of units selected in the stratum

and N is the size of the population of the stratum).

Imputation

16.70 Imputation involves supplying a value for a non-responding unit, or to replace 'suspect' data. Imputation methods fall into three groups:

- the imputed value may be derived from other information supplied by the respondent;
- the imputed value may be derived from information supplied by other similar respondents in the current survey; and
- the values supplied by the respondent in previous surveys may be modified to derive a value.

Three of the imputation methods used in labour-related surveys are described below.

16.71 **Deductive imputation** involves correcting a missing or erroneous value by using other information that reveals the correct answer. For example, a response of 18,000 has been given where respondents have been asked to reply in '$000s' and where the expected range of responses is 13-21. A quick examination of other parts of the form shows that $18,000 is very likely the amount actually spent by the respondent, so 18,000 is 'corrected' to 18.

16.72 **Central-value imputation** involves replacing a missing or erroneous item with a value considered to be 'typical' of the sample or sub-sample concerned. Live respondent mean is an example of central-value imputation. This technique involves calculating the average stratum value for the data item of interest across all responding live units in the stratum, and assigning this value to all live non-responding units in the stratum.

16.73 **Cold-deck imputation** involves using previous survey data to amend items which fail edits. It may involve copying data from the previous survey cycle to the current cycle. One specific example of this type of imputation is Beta imputation, which involves estimating missing values by applying an imputed growth rate to the most recently reported data for these units, provided that data have been reported in either of the two previous periods.

Adjustments for outliers

16.74 When adjusting for outliers, a compromise is always necessary between the variability and bias associated with an estimate. There are two methods available for dealing with outliers. Historically the ABS has used the 'surprise outlier' approach for most business surveys, but over time has gradually changed over to using 'winsorization'.

Surprise outlier approach

16.75 Generally, this technique is used to deal with a selected unit which is grossly extreme for a number of variables. The approach treats each outlier as if it were the only extreme unit in the stratum population. The outlier is given a weight of one, as if it had been selected in a CE stratum. As a result of the outlier's movement to the CE stratum, the weight for units in the outlier's selection stratum has to be recalculated, as the population and sample size have effectively been reduced by one. This has the effect that the other population units which would have been represented by the outlier are now represented by the average of the other units in the stratum. Therefore the choice of treatments for a suspected outlier using the surprise outlier approach are either for it to represent all of the units it would normally represent or to represent no units other than itself. It is preferable to set a maximum number of surprise outliers which can be identified in any one survey.

Winsorizing technique

16.76 This technique is a more flexible approach. Here a value is considered to be an outlier if it is greater than a predetermined cutoff. The effect of the outlier on the estimates is reduced by modifying its reported value.

16.77 On application of the winsorization formula, sample values greater than the cutoff are replaced by the cutoff plus a small additional amount. The additional amount is the difference between the sample value and the cutoff, multiplied by the stratum sampling fraction. Thus winsorization has most impact in strata with low sampling fractions, and the impact decreases as sampling fractions increase. Effectively, winsorization results in the outlier only representing itself, with the remaining population units that would have been represented by the outlier being instead represented by the cutoff.

TIME SERIES ESTIMATES

16.78 Time series are statistical records of various activities measured at more or less regular intervals of time, over relatively long periods. Data collected in irregular surveys do not form time series. The following section outlines the various elements of time series and outlines the ABS method of calculating seasonally adjusted and trend estimates.

16.79 ABS time series statistics are published in three forms: original, seasonally adjusted and trend.

16.80 **Original estimates** are the actual estimates the ABS derives from the survey data or other non-survey sources. Original estimates are composed of trend behaviour, systematic calendar related influences and irregular influences.

16.81 **Systematic calendar related influences** operate in a sustained and systematic manner that is calendar related. The two most common of these influences are seasonal influences and trading day influences.

16.82 **Seasonal influences** occur for a variety of reasons.

- They may simply be related to the seasons and related weather conditions such as warmth in summer and cold in winter. Weather conditions that are out of character for a particular season, such as snow in summer, would appear as irregular, not seasonal, influences.

- They may reflect traditional behaviour associated with various social events (e.g. Christmas and the associated holiday season).

- They may reflect the effects of administrative procedures (e.g. quarterly provisional tax payments and end of financial year activity).

16.83 **Trading day influences** refer to activity associated with the number and types of days in a particular month, as different days of the week often have different levels of activity. For instance, a calendar month typically comprises four weeks (28 days) plus an extra two or three days. If these extra days are associated with high activity, then activity for the month overall will tend to be higher.

16.84 Seasonal and trading day factors are estimates of the effect that the main systematic calendar related influences have on ABS time series. These evolve to reflect changes in seasonal and trading patterns of activity over the life of the time series, and are used to remove the effect of seasonal and trading day influences from the original estimates.

16.85 **Seasonally adjusted estimates** are derived by removing the systematic calendar related influences from the original estimates. Seasonally adjusted estimates capture trend behaviour, but still contain irregular influences that can mask the underlying month to month or quarter to quarter movement in a series. Seasonally adjusted estimates by themselves are only relevant for sub-annual collections.

16.86 **Irregular influences** are short term fluctuations which are unpredictable and hence are not systematic or calendar related. Examples of irregular influences are those caused by one-off effects such as major industrial disputes or abnormal weather patterns. Sampling and non-sampling errors that behave in an irregular or erratic fashion with no noticeable systematic pattern are also irregular influences.

16.87 **Trend estimates** are derived by removing irregular influences from the seasonally adjusted estimates. As they have neither systematic, calendar related influences nor irregular influences present in them, they are a measure of the underlying behaviour of the series.

CALCULATION OF TREND ESTIMATES

16.88 Trend estimates are produced by smoothing the seasonally adjusted series using a statistical procedure based on Henderson moving averages. At each survey cycle the trend estimates are calculated using a centred x-term Henderson moving average of the seasonally adjusted series. The moving averages are centred on the point in time at which the trend is being estimated. The number of terms used to calculate the trend estimates varies across surveys. Generally, ABS monthly surveys use a 13-term Henderson moving average and quarterly surveys use a 7-term Henderson moving average.

16.89 Estimates for the most recent survey cycles cannot be calculated using the centred moving average method as there are insufficient data to do so. Instead, alternative approaches that approximate the smoothing properties of the Henderson moving average are used - such as asymmetric averages. This can lead to revisions in the trend estimates for the most recent survey cycles until sufficient data are available to calculate the trend using the centred Henderson moving average. Revisions of trend estimates will also occur with revisions to the original data and re-estimation of seasonal adjustment factors.

RELIABILITY OF ESTIMATES

16.90 The accuracy of an estimate refers to how close that estimate is to the true population value. Where there is a discrepancy between the value of the sample estimate and the true population value, the difference between the two is referred to as the 'error of the sampling estimate'. The total error of the sampling estimate results from two types of error:

- sampling error - errors which occur because data were obtained from only a sample rather than the entire population; and
- non-sampling error - errors which occur at any stage of a survey and can also occur in censuses.

16.91 All ABS data are subject to one or both of these types of errors. The following section provides further information on both sampling and non-sampling error and describes various measures of each.

SAMPLING ERROR

16.92 Sampling error equals the difference between the estimate obtained from a particular sample and the value that would be obtained if the whole survey population were enumerated. It is important to consider sampling error when publishing survey results, as it gives an indication of the accuracy of the estimate and therefore reflects the importance that can be placed on interpretations. For a given estimator and sample design, the expected size of the sampling error is affected by how similar the units in the target population are, and the sample size.

Variance

16.93 Variance is a measure of sampling error that is defined as the average of the squares of the deviation of each possible estimate (based on all possible samples for the same design) from the expected value. It gives an indication of how accurate the survey estimate is likely to be, by measuring the spread of estimates around the expected value. For probability sampling, an estimate of the variance can be calculated from the data values in the particular sample that is generated.

16.94 Methods used to calculate estimates of variance in ABS labour-related surveys are outlined below.

- **Jack-knife** - this method starts by dividing the survey sample into a number of equally sized groups (replicate groups), containing one or more units. Pseudo-estimates of the population total are then calculated from the sample by excluding each replicate group in turn. The jack-knife variance is derived from the variation of the respective pseudo-estimates around the estimate based on the whole sample. This method is used in a number of labour-related business surveys.

- **Ultimate cluster variance** - this method is used in multi-stage sampling schemes (see previous explanation of multi-stage sampling), and involves using the variation in estimates derived from the first-stage units to estimate the variance of the total estimate. This method is used in the Survey of Employee Earnings and Hours.

- **Split halves** - this method involves dividing the sample into half and, from each half, obtaining an independent estimate of the total. The variance estimate is produced using the square of the difference of these estimates. Variations of the split halves method for calculating variance estimates are used in a number of household surveys including the Labour Force Survey.

16.95 The variances indicated in ABS household survey publications are generally based on models of each survey's variance. The variances for a range of estimates are calculated using one of the above methods and a curve fitted to the results. This curve indicates the level of variance which could be expected for a particular size of estimate.

Standard Error (SE)

16.96 The most commonly used measure of sampling error is called the standard error. The standard error is equal to the square root of the variance. An estimate of the standard error can be derived from either the population variance (if known) or the estimated variance from the sample units. Any estimate derived from a probability based sample survey has a standard error associated with it (called the standard error of the estimate). The main features of standard errors are set out below.

- Standard errors indicate how close survey estimates are likely to be to the expected population values that would be obtained from a census conducted under the same procedures and processes.
- Standard errors provide measures of variation in estimates obtained from all possible samples under a given design.
- Small standard errors indicate that variation in estimates from repeated samples is small, and that therefore it is likely that sample estimates will be close to the true population values, regardless of the sample selected.
- Estimates of standard errors can be obtained from any probability sample - different random samples will produce different estimates of standard errors.
- Standard errors calculated from survey samples are themselves estimates and thus also subject to sampling error.
- When comparing survey estimates, statements should be made about the standard errors of those estimates.
- Standard errors can be used to work out **confidence intervals**. This concept is explained below.

Confidence Interval (CI)

16.97 A confidence interval is defined as an interval, centred on the estimate, with a prescribed level of probability that it includes the true population value (if the estimator is unbiased) or the mean of the sampling distribution (if the estimator is biased). Estimates from ABS surveys are usually unbiased.

16.98 Estimates are often presented in terms of a confidence interval. Most commonly**,** confidence intervals are constructed for 68%, 95%, and 99% levels of probability. The true value is said to have a given probability of lying within the constructed interval. For example:

- 68% chance that the true value lies within 1 standard error of the estimate (2 chances in 3).
- 95% chance that the true value lies within 2 standard errors of the estimate (19 chances in 20).
- 99% chance that the true value lies within 3 standard errors of the estimate (99 chances in 100).

16.99 Confidence intervals are constructed using the standard error associated with an estimate. For example, a 95% confidence interval is equivalent to the survey estimate plus or minus two times the standard error of the estimate. Therefore, if the sample survey estimate of a variable was 100 and the estimate had a standard error of 10, the 95% confidence interval could be expressed: "we are 95% confident that the true value of the variable of interest lies within the interval [80, 120]".

Relative Standard Error (RSE)

16.100 Another measure of sampling error is the relative standard error (RSE). This is the standard error expressed as a percentage of the estimate. Since the standard error of an estimate is generally related to the size of the estimate, it is not possible to deduce the accuracy of the estimate from the standard error without also referring to the size of the estimate. The relative standard error avoids the need to refer to the estimate, since the standard error is expressed as a proportion of the estimate. RSEs are useful when comparing the variability of population estimates of different sizes. They are commonly expressed as percentages.

16.101 Very small estimates are subject to high RSEs which detract from their usefulness. In ABS labour-related statistical publications, estimates with an RSE greater than 25% but less than 50% have an asterisk (*) displayed beside the estimate, indicating they should be used with caution. Estimates with an RSE greater than 50% have two asterisks (**) displayed beside the estimate, indicating they are so unreliable as to detract seriously from their value for most reasonable uses.

NON-SAMPLING ERROR

16.102 Non-sampling error refers to all other errors in the estimate. Non-sampling error can be caused by non-response, badly designed questionnaires, respondent bias, interviewer bias, collection bias, frame deficiencies and processing errors. It is often difficult and expensive to quantify non-sampling error.

16.103 Non-sampling errors can occur at any stage of the process, and in both censuses and sample surveys. Non-sampling errors can be grouped into two main types: systematic and variable. Systematic error (called bias) makes survey results unrepresentative of the population value by systematically distorting the survey estimates. Variable error can distort the results on any given occasion, but tends to balance out on average over time.

Reducing non-sampling error

16.104 Every effort is made to minimise non-sampling error in ABS surveys at every stage of the survey, through careful design of collections, and the use of rigorous editing and quality control procedures in the compilation of data. Some of the approaches adopted are listed below.

- Reducing frame deficiencies - refer to paragraphs 16.32 to 16.38 above.

- Reducing non-response - non-response results in bias in the estimate because it is possible the non-respondents have different characteristics to respondents, leading to an under-representation of the characteristics of non-respondents in the sample survey estimate. The ABS pursues a policy of intensive follow up of non-respondents. This includes multiple visits or telephone calls in an attempt to contact respondents, and letters requesting compliance with the survey. Partial non-response is also followed up with respondents.

- Reducing instrument errors - these errors relate to poor questionnaire design, leading to questions which are not easily understood by respondents, and hence incorrect responses. This is particularly relevant for household surveys. The ABS ensures that all household survey questionnaires are carefully tested using cognitive testing, and dress rehearsals of the survey before it is officially conducted. New business survey questionnaires and additional questions in business surveys are also rigorously tested before they are introduced.

Measures of non-sampling error

16.105 Non-sampling error is difficult to quantify; however, an indication of the level of non-sampling error can be determined from a number of quality measures. These include:

- Response rates - the number of responding units in a survey expressed as a proportion of the total number of units selected (excluding deaths). Response rates can also be calculated for individual questions within a survey.
- Imputation rates - the number of responses which need to be imputed expressed as a proportion of the total number of responses.
- Coverage rates - an estimate of the proportion of units in the target population which are not covered by the frame.
- Any Responsible Adult rates - the number of responding units in a survey for which information was supplied by a responsible adult rather than personally, expressed as a proportion of the total number of responding units. Any Responsible Adult rates can only be calculated for household surveys. For further information on personal interview and Any Responsible Adult collection methodologies, see paragraphs 16.11 to 16.12.

OUTPUT

16.106 The ABS's objectives in dissemination are to ensure widespread availability of information, while recovering the marginal costs involved in providing products and services for private benefit.

16.107 To meet the ABS's 'public good' obligations, the main findings of statistical collections and statistical reports on matters of public interest are made available free of charge to the community via the media. ABS publications are made available free to parliamentarians, major news media organisations, and parliamentary, public and tertiary institution libraries. In addition, the ABS conducts a Library Extension Program within 515 libraries participating throughout Australia. These libraries are provided with free ABS publications and some electronic services to meet the needs of their local communities. Free access is also available to selected statistics on the ABS website (www.abs.gov.au).

16.108 The ABS policy of charging is intended to serve four main purposes:

- to enable the demand for ABS products and services to be used as a more reliable indicator of how ABS resources should be used;
- to encourage users to address their real needs for ABS products and services;
- to relieve the general taxpayer of those elements of the cost of the statistical services which have specific and identifiable value to particular users; and
- to promote sensible investment in client service facilities.

16.109 A number of international agencies, including the International Monetary Fund and the ILO, have put forward a range of proposals and guidelines for the dissemination of data including: the methodology of their collection and compilation, and evaluation as to their accuracy; relevance to the phenomena measured; and quality of the output. In particular, the ILO at its 1998 ICLS endorsed a set of 20 guidelines concerning dissemination practices for labour statistics (the ICLS Guidelines can be found on the ILO website at the following address: http://www.ilo.org/public/english/bureau/stat/standards/guidelines/index.htm). The ILO guidelines, and a comparison of these guidelines with ABS practice, are contained in the Appendix.

DISSEMINATION MEDIA

16.110 The ABS uses a range of media for the dissemination of labour statistics but, in line with clients' preferences, publications are the prime release medium and are available in both printed and electronic form.

16.111 In addition to publications, a range of other dissemination media are used in the release of labour statistics. The ABS produces, free of charge, a quick reference information service for basic statistical information, including information on labour statistics. The service operates in response to telephone calls, email, correspondence and personal visits. The ABS also offers information consultancy services on a fee for service basis, for clients requesting more complex information.

16.112 Confidentialised Unit Record Files are available for some labour collections. These files contain the responses received for each unit in the survey, with any identifying information removed.

16.113 The ABS offers a range of subscription services including AusStats and ABS@. AusStats and ABS@ are both web-based information services making the ABS standard product range available on-line. Information available through AusStats and ABS@ includes: all ABS publications from 1998 onwards in Adobe Acrobat format (.pdf); multi-dimensional datasets in SuperTABLE format; Census Basic Community Profiles to the Statistical Local Area level in Excel spreadsheet format; and a range of free summary information including Main Features, Release Advices and Australia Now. AusStats is accessed through the ABS web site and offers a number of subscription plans to suit different requirements. The ABS@ service, which is replicated daily onto the Intranets of subscribing organisations, enables all staff within those organisations to access ABS services.

CONFIDENTIALITY

16.114 All releases of data from the ABS are confidentialised to ensure that no unit (e.g. person or business) is able to be identified. The ABS applies a set of rules, concerning the minimum number of responses required to contribute to each data cell of a table, and the maximum proportion that any one respondent can contribute to a table cell, to ensure that information about specific units cannot be derived from published survey results.

16.115 In some instances it is not possible to confidentialise responses from businesses that contribute substantially to a data cell. In this case, agreement is sought from the business for their data to still be published. If agreement is not reached, all affected data cells are suppressed.

DATA COMPARABILITY OVER TIME

16.116 The ABS aims to produce consistent and comparable time series of data by minimising changes to ongoing surveys. However, the frequency of collection, collection and sample methods, concepts, data item definitions, classifications and time series analysis techniques are all subject to maintenance, change and/or development.

16.117 The desire for comparable data must be balanced with a requirement for data to remain relevant. In addition, sound survey practice requires careful and continuing maintenance and development to ensure the integrity of the data and the efficiency of the collection. Some survey features are reviewed regularly, while others are changed only as the need arises. For example, the sample design for the Labour Force Survey is based on the Population Census (conducted every five years), and is therefore reviewed on a five-yearly cycle. Updates to the seasonally adjusted and trend series resulting from time series analysis are also changed regularly.

16.118 On the other hand, irregular changes to questionnaires may arise from:

- changes in international recommendations (these usually occur infrequently);
- changes in local needs or conditions;
- reviews of ABS data standards, such as changes to the Industry and Occupation classifications;
- changes to population frames, such as the Business Register; and
- developments in ABS collection methods, such as the introduction of telephone interviewing or computer assisted personal interviewing.

16.119 Changes to ABS surveys which affect the comparability of data over time are usually documented in the explanatory notes of survey publications. Changes to individual labour-related surveys which have occurred to date are also summarised in subsequent chapters.

Previous Page Next Page

This page last updated 9 February 2006